DOCUMENT RESUME

ED 054 815 LI 003 091

AUTHOR Lay, W. Michael; Petrarca, Anthony E.

TITLE Modified Double KWIC Coordinate Index. Refinements in Main Term and Subordinate Term Selection.

INSTITUTION Ohio State Univ., Columbus. Computer and Information Science Research Center.

SPONS AGENCY National Science Foundation, Washington, D.C. Office of Science Information Services.

PUB DATE 70

NOTE 26p.; (4 References)

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS Automation; Computer Programs; *Coordinate Indexes; *Indexes (Locaters); Indexing; *Information Retrieval; *Information Storage; *Permuted Indexes; Statistics

IDENTIFIERS *Key Word in Context; KWIC; Ohio State University

ABSTRACT

A new approach to creating the double KWIC Coordinate Index consists of extracting potential main terms directly from the titles (or title-like Phrases) instead of from a KWIC index of the titles, and sorting and temporarily retaining the potential index entries in a KWOC type format until other conditions are examined. After all of the titles have been processed and the actual main terms have been selected, if the number of titles containing a particular main term exceeds an abitrarily assigned threshold value, conventional double KWIC (permuted) subordinate entries are created. If the threshold value is not exceeded, KWOC-type (non-permuted) subordinate entries are created. Other refinements which improve the overall quality of the index have also been introduced. In addition, the new approach facilitates gathering of statistics which will help to reduce the manual and intellectual effort previously required for selection of main terms. The system design for creating the modified Double KWIC Coordinate Index is discussed together with some operating statistics for producing such an index. The use of statistical relationships to create cross references between certain types of terms is also explored. (Author/NH)

ED054815

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Technical Report Series
(OSU-CISRC-TR-70-10)

MODIFIED DOUBLE KWIC COORDINATE INDEX.

REFINEMENTS IN MAIN TERM AND SUBORDINATE TERM SELECTION

by

W. Michael Lay and Anthony E. Petrarca

Work Performed Under

Grant No. 534.1, National Science Foundation

Computer and Information Science Research Center
The Ohio State University
Columbus, Ohio 43210

1970

LI 003 091

PREFACE

This report is the result of research on information storage and retrieval supported in part by Grant Number GN-534.1 from the Office of Science Information Service, National Science Foundation to the Computer and Information Science Research Center, The Ohio State University.

An earlier version of this paper was presented at the 7th Annual National Information Retrieval Colloguium (ANIRC), Philadelphia, Pennsylvania in May, 1970 and it will be published in the Proceedings of that meeting.

The Computer and Information Science Research Center of The Ohio State University is an interdisciplinary research organization which consists of the staff, graduate students, and faculty of many University departments and laboratories. This report is based on research performed in the Department of Computer and Information Science.

The research was administered and monitored by The Ohio State University Research Foundation as Project RF 2218.

ABSTRACT

Our prototype model of the double KWIC Coordinate Index suggested that the overall quality of such an index could be significantly improved by modification of the techniques for creation of the index entries. In the new approach, potential main terms are extracted directly from the titles (or title-like phrases) instead of from a KWIC index of the titles, and the potential index entries so created are sorted and temporarily retained in a KWOC-type format until other conditions are examined. After all of the titles have been processed and the actual main terms have been selected, if the number of titles containing a particular main term exceeds an arbitrarily assigned threshold value, conventional double KWIC (permuted) subordinate entries are created. If the threshold value is not exceeded, KWOC-type (non-permuted) subordinate entries are created. Other refinements which improve the overall quality of the index have also been introduced. In addition, the new approach facilitates gathering of statistics which will help to reduce the manual and intellectual effort previously required for selection of main terms. The system design for creating the modified Double KWIC Coordinate Index is discussed together with some operating statistics for producing such an index. The use of statistical relationships to create cross references between certain types of terms is also explored.

TABLE OF CONTENTS

	Page
Preface	ii
ABSTRACT	iii
INTRODUCTION	1
INITIAL SYSTEM DESIGN	2
SYSTEM DESIGN FOR THE NEW APPROACH	4
INDEX CONSTRUCTION VARIABLES	7
HUMAN INTERFACE REQUIREMENTS	8
CONCLUSION	10
ACKNOWLEDGMENTS	10
REFERENCES	11
FIGURES	

Introduction

In several earlier papers [1,2], we described the general characteristics of a new method for generating high-quality printed indexes by relatively simple automated techniques. The technique, which is based on an extension of the KWIC indexing concept, leads to the production of a new type of index called the double KWIC coordinate index. This index has many of the qualities of an articulated subject index, but it can be produced more simply and less expensively by automated techniques because it does not require the depth of syntactic analysis usually required for the automatic production of articulated subject indexes.

For our initial model, the double KWIC coordinate index was constructed by applying two fundamental operations to title or title-like phrases; (Figure 1)

1. The first significant word (or phrase) of the title was extracted as a potential main term and replaced by an asterisk to indicate its position within the title;
2. The remaining words of the title were rotated so as to allow each significant word to appear as the first word of a wrap-around subordinate entry under the potential main term.

These operations were applied to each significant word in the title and for all titles in the source. The entries so created were sorted, first on the main terms, then on the basis of the words in the subordinate entry. Each subordinate entry was printed with as much of the surrounding context as one line would allow.

Our first application of the double KWIC coordinate indexing technique was for construction of a prototype index, a portion of which is displayed in Figure 2. The prototype not only provided us with a model to illustrate

the potential advantages of this new automatic indexing technique, but also suggested a number of ways in which the model could be refined. One immediately obvious refinement pertains to the often encountered situation illustrated in Figure 3 where the permuted subordinate entries under the main term were all derived from the same title. Obviously, there is little justification for permuting subordinate entries in situations like this. Another observation is illustrated in Figure 4 for those cases where an indexable word or phrase occurs more than once in a title. The title from which these particular entries were created contained the word "INDEX" twice. For each occurrence, it was extracted as if it were a different main term. Subsequent rotations of the remaining words in the title produced pairs of nearly identical subordinate entries in the resulting index. Observations such as those just described caused us to reexamine the approach used to construct our prototype model.

Initial System Design

The system designed for creation of the prototype double KWIC coordinate index (Figure 5) was as follows. The first phase required generation of KWIC index records from the source titles. Since all of the words appearing in the index column of the conventional KWIC index would become candidates for potential main terms in the double KWIC coordinate index, a main term stoplist was invoked in the KWIC index step to preclude creation of index entries for all words of low index value which we did not want to appear as the first word of a main index term in the double KWIC coordinate index. Potential main terms for the double KWIC index were generated in the second step by extracting individual keywords or phrases (word strings) from the index column of the KWIC index. After each potential main term was extracted, the remaining portion of the title was rotated so as to create permuted

subordinate entries. A second stoplist consisting primarily of articles, prepositions, and conjunctions precluded generation of subordinate entries beginning with words appearing in this stoplist.

The algorithm for generating potential main terms defined a word as a string of characters bounded by spaces. A potential main term could consist of a single word or a set of contiguous words up to some specified upper limit. If a punctuation mark occurred at the end of any word, it was removed during creation of a potential main term. Also, word strings for which the last word of the string was on the subordinate stoplist were not generated as potential main terms.

The index records generated by the above procedures were sorted first on the basis of potential main terms and then on the basis of the words in the subordinate entry. From this sorted file, a printed list of all potential main terms generated by the procedure was obtained so that the indexer could choose the actual main terms which would appear in the final index. These selections were made during the final print phase via sequence numbers assigned to the potential main terms in the printed list.

The problems mentioned earlier for our prototype index obviously resulted from too close adherance to the principles of KWIC index construction. Once a potential main term was extracted from a title the remaining portion of the title was always permuted regardless of whether the potential main term occurred more than once in a given title or whether it occurred only once in the entire set of titles being indexed. Fully rotated subordinate entries were constructed for all potential main terms whether or not they were selected for inclusion in the final index. This indiscriminate approach to permuted subordinate entry construction not only created the problems mentioned

before (Figure 3 and Figure 4), but also needlessly increased the cost of constructing the index. Although some of these problems had been anticipated beforehand, we decided to generate all second order permutations of the titles for the prototype index on the premise that word and phrase patterns generated by these permutations might provide some insight into the problems of main term and subordinate term selection.

System Design for the New Approach

To overcome many of the problems encountered in the prototype model, a slightly different approach for construction of the double KWIC coordinate index was used (Figure 6). The basic differences between the old and new approach are: (1) the potential main terms are now extracted directly from the titles (or title-like phrases) instead of from a KWIC index of the titles, and the potential index entries so created are temporarily retained in a KWOC-type format until other conditions are examined; (2) after all of the titles have been processed and the actual main terms have been selected, if the number of titles containing a particular main term exceeds an arbitrarily assigned threshold value, conventional double KWIC (permuted) subordinate entries are created; if the threshold value is not exceeded KWOC-type (non-permuted) subordinate entries are created.

The system design for the new approach is illustrated in Figure 7. The first phase consists of the extraction of all potential main terms from the titles being indexed, word significance still being based on appropriate stoplists. The new algorithm for generating potential main terms defines a word as a string of characters bounded by a set of delimiters. These delimiters are partitioned into two classes, terminal and non-terminal (Figure 8), and the function of each is described below.

A potential main term may consist of a single word or a set of contiguous

words up to some specified upper limit; it must have the following three attributes:

1) the first word of the potential main term must not be on the main term stoplist;

2) the last word of a condidate contiguous set must not be on the subordinate stoplist;

3) all words in a candidate contiguous set must be separated by non-terminal word delimiters.

The first and second attributes are the same as those previously required in the KWIC indexing and potential main term generation phases, respectively, of the earlier approach. Because certain punctuation marks between words usually signify introduction of a new concept, requirement of the third attribute was introduced to assure generation of potential main terms describing only a single concept. Finally the new approach identifies multiple occurrences of a potential main term in any particular title being indexed; hence, only unique potential main terms are generated from a given title. Figure 8 illustrates the potential main terms that would be generated from a title on the basis of the above criteria. For one of the potential index entries illustrated therein, note the treatment for multiple occurrences of a main term (DATA) in a title. This treatment precluded the possibility of generating groups of nearly identical subordinate entries such as encountered in the initial model (Figure 4).

After all the titles from a given source have been processed and the potential index entries have been sorted, a printed list of all potential main terms, referenced by sequence number, is prepared. This list also

includes frequency data for the number of titles in which that particular main term occurs. Figure 9 illustrates some potential-main-term listings from a recent production run.

At this point, a human interface step is required for selection of the actual main terms which are to appear in the modified double KWIC coordinate index. The sequence numbers for the desired main terms (e.g. #29 and #33 in Figure 9) and the threshold value for controlling the relative number of permuted and non-permuted subordinate entries are supplied as input parameters to the next processing program. The actual entries for the index are then sorted and printed in accordance with any previously supplied display specifications (see Figure 10). An additional display feature for permuted subordinate entries under the new approach enables one to more easily identify certain proximity relationships (and hence, semantic relationships) between main terms and subordinate terms. This is accomplished by displaying the replacement asterisk for the main term in the left hand margin of the subordinate entry when the main term immediately precedes the first word of the wrap-around entry.

The new systems design enables one to produce a range of index types which can vary in size, quality, and cost. This can be accomplished simply by varying the threshold value which controls the relative number of permuted and non-permuted subordinate entries. For example by using a threshold value of zero one can produce an index in which all of the subordinate entries are permuted as was exemplified by our initial prototype index. On the other hand by using an extraordinarily high threshold value one can produce a straight KWOC index. In between these two extremes one can produce indexes with

varying combinations of both types of entries. The advantage of using KWOC-type subordinate entries is that they result in a considerable reduction in the size of the printed index. But, as the number of KWOC-type subordinate entries under a main term is increased, the accessibility to subordinate concepts described therein is significantly reduced and the advantages of double KWIC coordinate indexing are lost. However, if one is willing to concede that accessibility to subordinate concepts is not significantly reduced when the number of KWOC-type subordinate entries is small, one can achieve a significant reduction in the over all size and cost of the printed index by using a low threshold value for controlling the generation of permuted and non-permuted subordinate entries. For example, the size of our prototype index was reduced by 40% simply by using a threshold value of one.

Index Construction Variables

Generally speaking, the cost factors involved in producing a double KWIC coordinate index vary considerably with respect to different data bases. Among the more important parameters are (1) the number of titles to be processed, (2) the average number of significant words per title, (3) the average number of main terms generated per title, and (4) the type of subordinate entries created (i.e. permuted or non-permuted). We have found in many instances that comparable numbers of documents taken from different disciplines show a wide range in the average number of significant words per title and a corresponding wide range in the average number of main terms per title. Even then, the percentage of actual main terms finally chosen for the index is highly dependent on the data being processed and the decisions of the analyst choosing the main terms.

Statistics given below for some indexes produced thus far illustrate the increased efficiency of the new system design over the old, and also illustrate the effect of changing one of the parameters mentioned above. Both programs were written in PL/1 and executed on an IBM 360/75 computer operating under OS 360/MVT. All comparisons were made on the input data used for construction of the prototype index which consisted of 90 titles containing an average of approximately four significant words per title. Using the earlier approach, about two minutes of computer time were required to construct the prototype index. The same index was created in less than 45 seconds by the modified technique, using a threshold of zero. By using a threshold value of one, the processing time was reduced to 30 seconds. Under these last conditions, a modified double KWIC coordinate index containing some KWOC-type subordinate entries was produced; as mentioned earlier, the size of the printed index was also reduced by about 40% with this change. We are presently experimenting with larger data bases to see how some of the parameters are affected by the size of the file.

Human Interface Requirements

At present, creation of double KWIC coordinate indexes requires two human-interface steps after the input data has been converted into machine readable form. The first is to determine the words which have to appear on the stoplists. For a new subject area, this step will usually require production of a trial index to determine what words must be included on these lists. Once the stoplists have been created for a particular subject area they can be used repeatedly with only periodic updating.

A second human interface is required for selection of the actual main terms which are to appear in the final index. As mentioned earlier, these are chosen from the list of potential main terms generated from the input data.

We are currently investigating several possible methods of automating this step. One approach is to use the frequency data for the number of titles containing a potential main term (Figure 9). The potential-main-term list represents, mathematically, a collection of trees in leftlist form. Each of the trees is rooted in one-word potential main terms as depicted in Figure 11. The level in the tree of any potential main term can be determined by the number of words in that potential term. The problem, then, involves development of appropriate algorithms to select a set of meaningful index terms from this tree to cover all of the titles from which the tree was generated.

A second approach would be to generate only one-word main terms as is presently done in other automatically produced indexes (e.g., Computing Reviews [3], Permuterm [4]). This would entirely eliminate the need for the second human interface step.

By taking advantage of the tree structure, better index terms could be chosen. For example, using the tree structure illustrated in Figure 11, DATA RETRIEVAL SYSTEM could be chosen over the parent term DATA RETRIEVAL since each occurs the same number of times in the titles being indexed. To eliminate redundancy in the index, the articles posted under DATA RETRIEVAL SYSTEM would not be posted under the term DATA since the more generic heading would be used only for articles which are not indexed under a more specific term in the tree structure. Cross referencing between the two could be provided, if desirable, by automatic creation of "see" and "see also" references.

The liklihood of being able to totally automate the selection process without including some occasional "weird" main terms in the index is precluded by the fact that any algorithm based soley on the level frequency-data of the tree fails when each level of the tree has a frequency of occurrence of one.

For an on-going indexing service, a possible solution to this problem would be to use the human interface approach initially to build the vocabulary for an authority list. Once the vocabulary has reached a certain size the authority list would be used to make automatic selections similar to those made by a human analyst as discussed earlier. For terms which are new to the vocabulary, a priori selections could be made by the statistical approach, and these could be validated or altered a posteriori for entry into the authority list by a human analyst.

Conclusion

As has been shown the modified technique for construction of the double KWIC coordinate index not only is more efficient from an operational viewpoint, but also offers the added attraction of enabling one to produce a range of index types varying from a straight KWOC to a straight double-KWIC. Current research shows promise of being able to reduce the amount of human interface presently required for the production of these indexes.

Acknowledgments

Partial support of this work by the National Science Foundation, Office of Science Information Services, under Grant GN-534 is gratefully acknowledged. The authors also wish to thank the Office of Education for a fellowship to W.M.L., and the Ohio State University Instructional and Research Computer Center for underwriting the cost of much of the computer time required to carry out portions of this work.

References

1. Petrarca, A.E. and W.M. Lay, "The Double KWIC Coordinate Index. A New Approach for Preparation of High-Quality Printed Indexes by Automated Indexing Techniques", *J. Chem. Doc.*, 9, 256(1969).

2. Petrarca, A.E. and W.M. Lay, "The Double KWIC Coordinate Index II. Use of an Automatically Generated Authority List to Eliminate Scattering Caused by Some Singular and Plural Main Index Terms", *ASIS Proceedings*, 6, 277(1969).

3. *Comprehensive Bibliography of Computing Literature, Permuted and Subject Index to Computing reviews*, Association for Computing Machinery, New York, New York.

4. *Permuterm Subject Index*, Institute for Scientific Information, Philadelphia, Pennsylvania.

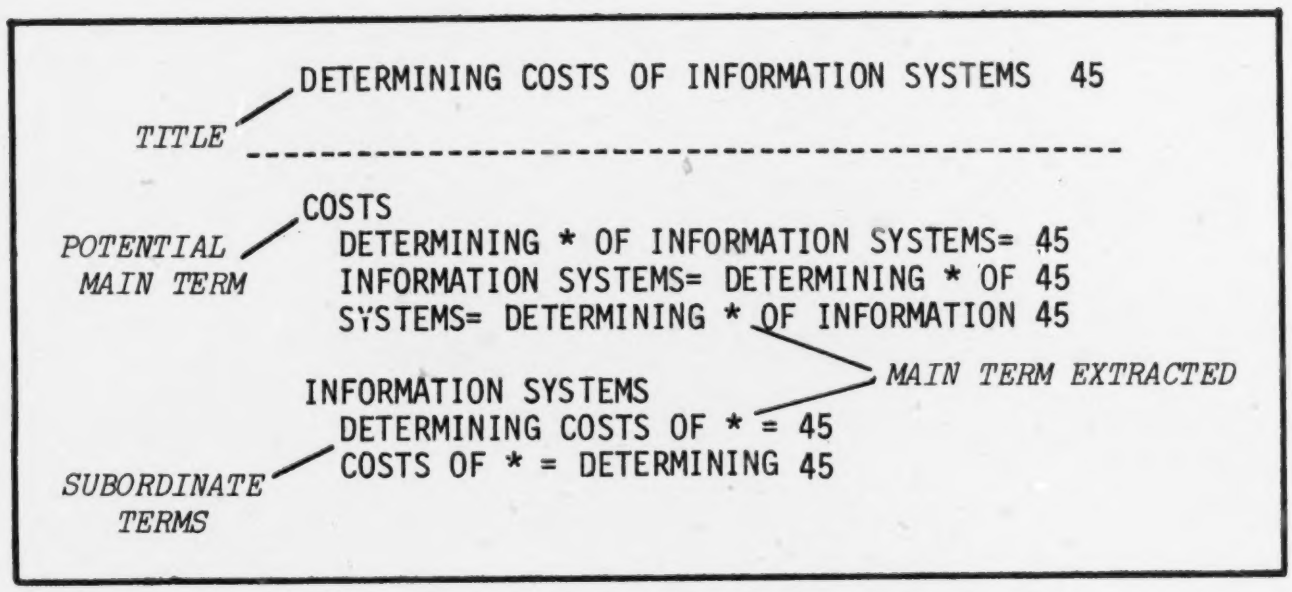
DETERMINING COSTS OF INFORMATION SYSTEMS 45
TITLE
COSTS
POTENTIAL MAIN TERM
DETERMINING * OF INFORMATION SYSTEMS= 45
INFORMATION SYSTEMS= DETERMINING * OF 45
SYSTEMS= DETERMINING * OF INFORMATION 45
MAIN TERM EXTRACTED
INFORMATION SYSTEMS
DETERMINING COSTS OF * = 45
SUBORDINATE TERMS
COSTS OF * = DETERMINING 45

Figure 1. Construction of potential entries for the prototype double KWIC coordinate index.

HISTORICAL BACKGROUND (FONT)
DATA COMPILING ACTIVITIES *..... * OF 3
HISTORY
BOOK_REVIEW: PROJECT * RETRIEVAL *..... B2-2
EDITORIAL: A SENSE OF * *..... F185
PROJECT * RETRIEVAL *..... BOOK_REVIEW: B2-2
RETRIEVAL *..... BOOK_REVIEW: PROJECT * B2-2
SENSE OF * *..... EDITORIAL: A F185
INDEX
ASSISTED PRIMARY * PREPARATION *..... COMPUTER 223
BIBLIOGRAPHIC, AND CATALOG ENTRIES **EW: COMPUTER FILING OF *, B2-2
BOOK_REVIEW: COMPUTER FILING OF *, BIBLIOGRAPHIC, AND CATALOG * B2-2
BOOK_REVIEW: MOSSBAUER EFFECT DATA * *..... B182
BOOK_REVIEW: REACTION * OF ORGANIC SYNTHESES *..... B257
BOOK_REVIEW: SADTLER STANDARD SPECTRA, NMR * *..... B182
CATALOG ENTRIES **EW: COMPUTER FILING OF *, BIBLIOGRAPHIC, AND B2-2
COMPUTER ASSISTED PRIMARY * PREPARATION *..... 223
COMPUTER FILING OF *, BIBLIOGRAPHIC, AND CATALOG ENTRIES **EW: B2-2
COMPUTER PROGRAM FOR AN INVERTED COORDINATE * ON MAGNETIC TAPE* 49
COORDINATE * ON MAGNETIC TAPE **MPUTER PROGRAM FOR AN INVERTED 49
DATA * *..... BOOK_REVIEW: MOSSBAUER EFFECT B182
EFFECT DATA * *..... BOOK_REVIEW: MOSSBAUER B182
ENTRIES **EW: COMPUTER FILING OF *, BIBLIOGRAPHIC, AND CATALOG B2-2
FILING OF *, BIBLIOGRAPHIC, AND CATALOG ENTRIES **EW: COMPUTER B2-2
INVERTED COORDINATE * ON MAGNETIC TAPE **MPUTER PROGRAM FOR AN 49
MAGNETIC TAPE **MPUTER PROGRAM FOR AN INVERTED COORDINATE * ON 49
MOSSBAUER EFFECT DATA * *..... BOOK_REVIEW: B182
NMR * *..... BOOK_REVIEW: SADTLER STANDARD SPECTRA, B182
ORGANIC SYNTHESES *..... BOOK_REVIEW: REACTION * OF B257
PREPARATION *..... COMPUTER ASSISTED PRIMARY * 223
PRIMARY * PREPARATION *..... COMPUTER ASSISTED 223
PROGRAM FOR AN INVERTED COORDINATE * ON MAGNETIC TAPE **MPUTER 49
REACTION * OF ORGANIC SYNTHESES *..... BOOK_REVIEW: B257
SADTLER STANDARD SPECTRA, NMR * *..... BOOK_REVIEW: B182
SEARCH: A COMPUTER PROGRAM FOR AN INVERTED COORDINATE * ON MAG* 49
SPECTRA, NMR * *..... BOOK_REVIEW: SADTLER STANDARD B182
STANDARD SPECTRA, NMR * *..... BOOK_REVIEW: SADTLER B182
SYNTHESES *..... BOOK_REVIEW: REACTION * OF ORGANIC B257
TAPE **MPUTER PROGRAM FOR AN INVERTED COORDINATE * ON MAGNETIC 49
TERM SEARCH: A COMPUTER PROGRAM FOR AN INVERTED COORDINATE * O* 49
WEIGHTED TERM SEARCH: A COMPUTER PROGRAM FOR AN INVERTED COORD* 49
INDEXES
APPROACH TO AUTOMATED VOCABULARY CONTROL IN * OF ORGANIC COMPO* 131
ARTICULATION IN THE GENERATION OF SUBJECT * BY COMPUTER *..... 170
AUTOMATED VOCABULARY CONTROL IN * OF ORGANIC COMPOUNDS **CH TO 131
COMPOUNDS **CH TO AUTOMATED VOCABULARY CONTROL IN * OF ORGANIC 131
COMPUTER *..... ARTICULATION IN THE GENERATION OF SUBJECT * BY 170
CONTROL IN * OF ORGANIC COMPOUNDS **CH TO AUTOMATED VOCABULARY 131
GENERATION OF SUBJECT * BY COMPUTER *..... ARTICULATION IN THE 170
ORGANIC COMPOUNDS **CH TO AUTOMATED VOCABULARY CONTROL IN * OF 131
SUBJECT * BY COMPUTER *..... ARTICULATION IN THE GENERATION OF 170
VOCABULARY CONTROL IN * OF ORGANIC COMPOUNDS **CH TO AUTOMATED 131
INDEXING
AIDS IN THE EVALUATION OF * TERMINOLOGY *..... COMPUTER 237
COMPUTER AIDS IN THE EVALUATION OF * TERMINOLOGY *..... 237
EVALUATION OF * TERMINOLOGY *..... COMPUTER AIDS IN THE 237
FORMULA *..... * POLYMERS BY 244
POLYMERS BY FORMULA *..... * 244
TERMINOLOGY *..... COMPUTER AIDS IN THE EVALUATION OF * 237
INDIVIDUAL USE
BIOMEDICAL INFORMATION RETRIEVAL: A COMPUTER-BASED SYSTEM FOR * 98
COMPUTER-BASED SYSTEM FOR * **MEDICAL INFORMATION RETRIEVAL: A 98
INFORMATION RETRIEVAL: A COMPUTER-BASED SYSTEM FOR * **MEDICAL 98
RETRIEVAL: A COMPUTER-BASED SYSTEM FOR * **MEDICAL INFORMATION 98
SYSTEM FOR * **MEDICAL INFORMATION RETRIEVAL: A COMPUTER-BASED 98
INDUSTRIAL CHEMICAL ANALYSIS
BOOK_REVIEW: ENCYCLOPEDIA OF *, VOL 4 *..... B258
BOOK_REVIEW: ENCYCLOPEDIA OF *, VOLS 2 AND 3 *..... B 59
ENCYCLOPEDIA OF *, VOLS 2 AND 3 *..... BOOK_REVIEW: B 59
ENCYCLOPEDIA OF *, VOL 4 *..... BOOK_REVIEW: B258
INFORMATION ANNOUNCEMENT SYSTEMS
COMMUNITY OF USERS *..... SELECTIVE * FOR A LARGE 142
LARGE COMMUNITY OF USERS *..... SELECTIVE * FOR A 142
SELECTIVE * FOR A LARGE COMMUNITY OF USERS *..... 142
USERS *..... SELECTIVE * FOR A LARGE COMMUNITY OF 142
INFORMATION CENTER ADMINISTRATION
BOOK_REVIEW: TECHNICAL *, VOL 3 *..... B257
TECHNICAL *, VOL 3 *..... BOOK_REVIEW: B257
INFORMATION DISTRIBUTION
AUTOMATIC * USING COMPUTER-COMPILED THESAURUS AND DUAL DICTION* 124
B.F. GOODRICH INFORMATION RETRIEVAL SYSTEM AND AUTOMATIC * USI* 124
COMPUTER-COMPILED THESAURUS AND DUAL DICTIONARY **ATIC * USING 124
DICTIONARY **ATIC * USING COMPUTER-COMPILED THESAURUS AND DUAL 124
DUAL DICTIONARY **ATIC * USING COMPUTER-COMPILED THESAURUS AND 124
GOODRICH INFORMATION RETRIEVAL SYSTEM AND AUTOMATIC * USING CO* 124
INFORMATION RETRIEVAL SYSTEM AND AUTOMATIC * USING COMPUTER-CO* 124
RETRIEVAL SYSTEM AND AUTOMATIC * USING COMPUTER-COMPILED THESA* 124
SYSTEM AND AUTOMATIC * USING COMPUTER-COMPILED THESAURUS AND D* 124
THESAURUS AND DUAL DICTIONARY **ATIC * USING COMPUTER-COMPILED 124
USING COMPUTER-COMPILED THESAURUS AND DUAL DICTIONARY **ATIC * 124
INFORMATION GROUPS
ADMINISTRATION OF TECHNICAL *, INTRODUCTORY REMARKS **OSIUM ON 110
INTRODUCTORY REMARKS **OSIUM ON ADMINISTRATION OF TECHNICAL *, 110
REMARKS **OSIUM ON ADMINISTRATION OF TECHNICAL *, INTRODUCTORY 110
SYMPOSIUM ON ADMINISTRATION OF TECHNICAL *, INTRODUCTORY REMAR* 110
TECHNICAL *, INTRODUCTORY REMARKS **OSIUM ON ADMINISTRATION OF 110
INFORMATION MANAGEMENT
BOOK_REVIEW: * IN ENGINEERING EDUCATION *..... B2-2
EDUCATION *..... BOOK_REVIEW: * IN ENGINEERING B2-2
ENGINEERING EDUCATION *..... BOOK_REVIEW: * IN B2-2
INFORMATION PROGRAM
BUILDING AN OPERATIONAL * *..... FACTORS IN 107
FACTORS IN BUILDING AN OPERATIONAL * *..... 107
OPERATIONAL * *..... FACTORS IN BUILDING AN 107
INFORMATION RETRIEVAL
AUTOMATIC INFORMATION DISTRIBUTION USING COMPUTER-COMPILED THE* 124
B.F. GOODRICH * SYSTEM AND AUTOMATIC INFORMATION DISTRIBUTION * 124
BIOMEDICAL *: A COMPUTER-BASED SYSTEM FOR INDIVIDUAL USE *.... 98
COMPUTER-BASED SYSTEM FOR INDIVIDUAL USE *.... BIOMEDICAL *: A 98
COMPUTER-COMPILED THESAURUS AND DUAL DICTIONARY **RUTION USING 124
DICTIONARY **RUTION USING COMPUTER-COMPILED THESAURUS AND DUAL 124
DISTRIBUTION USING COMPUTER-COMPILED THESAURUS AND DUAL DICTIO* 124
DUAL DICTIONARY **RUTION USING COMPUTER-COMPILED THESAURUS AND 124
GOODRICH * SYSTEM AND AUTOMATIC INFORMATION DISTRIBUTION USING* 124
INDIVIDUAL USE *.... BIOMEDICAL *: A COMPUTER-BASED SYSTEM FOR 98
INFORMATION DISTRIBUTION USING COMPUTER-COMPILED THESAURUS AND* 124
SYSTEM AND AUTOMATIC INFORMATION DISTRIBUTION USING COMPUTER-C* 124
SYSTEM FOR INDIVIDUAL USE *.... BIOMEDICAL *: A COMPUTER-BASED 98
THESAURUS AND DUAL DICTIONARY **RUTION USING COMPUTER-COMPILED 124
USE *.... BIOMEDICAL *: A COMPUTER-BASED SYSTEM FOR INDIVIDUAL 98
USING COMPUTER-COMPILED THESAURUS AND DUAL DICTIONARY **RUTION 124
INFORMATION SCIENCE
ANNUAL REVIEW OF * AND TECHNOLOGY *..... BOOK_REVIEW: B3-2
BOOK_REVIEW: ANNUAL REVIEW OF * AND TECHNOLOGY *..... B3-2
REVIEW OF * AND TECHNOLOGY *..... BOOK_REVIEW: ANNUAL B3-2
TECHNOLOGY *..... BOOK_REVIEW: ANNUAL REVIEW OF * AND B3-2

INFORMATION SCIENTISTS
ACADEMIC TRAINING PROGRAMS FOR * *..... SALARIES AND 118
PROGRAMS FOR * *..... SALARIES AND ACADEMIC TRAINING 118
SALARIES AND ACADEMIC TRAINING PROGRAMS FOR * *..... 118
TRAINING PROGRAMS FOR * *..... SALARIES AND ACADEMIC 118
INFORMATION SERVICES
CONTINUING EDUCATION IN TECHNICAL * *..... 115
COORDINATION AND INTEGRATION OF TECHNICAL * *..... 111
EDUCATION IN TECHNICAL * *..... CONTINUING 115
INTEGRATION OF TECHNICAL * *..... COORDINATION AND 111
TECHNICAL * *..... CONTINUING EDUCATION IN 115
TECHNICAL * *..... COORDINATION AND INTEGRATION OF 111
INFORMATION STORAGE
CHEMICALLY ORIENTED * AND RETRIEVAL SYSTEM, I. STORAGE AND VER* 43
INFORMATION **YSTEM, I. STORAGE AND VERIFICATION OF STRUCTURAL 43
ORIENTED * AND RETRIEVAL SYSTEM, I. STORAGE AND VERIFICATION O* 43
RETRIEVAL SYSTEM, I. STORAGE AND VERIFICATION OF STRUCTURAL IN* 43
STORAGE AND VERIFICATION OF STRUCTURAL INFORMATION **YSTEM, I. 43
STRUCTURAL INFORMATION **YSTEM, I. STORAGE AND VERIFICATION OF 43
SYSTEM, I. STORAGE AND VERIFICATION OF STRUCTURAL INFORMATION * 43
VERIFICATION OF STRUCTURAL INFORMATION **YSTEM, I. STORAGE AND 43
INFORMATION SYSTEM
CHEMICAL * *..... USE OF A NONUNIQUE NOTATION IN A LARGE-SCALE 192
EDITORIAL: A NATIONAL * *..... F 61
LARGE-SCALE CHEMICAL * *..... USE OF A NONUNIQUE NOTATION IN A 192
NATIONAL * *..... EDITORIAL: A F 61
NONUNIQUE NOTATION IN A LARGE-SCALE CHEMICAL * *..... USE OF A 192
NOTATION IN A LARGE-SCALE CHEMICAL * *..... USE OF A NONUNIQUE 192
USE OF A NONUNIQUE NOTATION IN A LARGE-SCALE CHEMICAL * *..... 192
INFORMATION SYSTEMS
BOOK_REVIEW: NONCONVENTIONAL SCIENTIFIC AND TECHNICAL * IN CUR* B3-2
COSTS OF * *..... DETERMINING 101
CURRENT USE **W: NONCONVENTIONAL SCIENTIFIC AND TECHNICAL * IN B3-2
DETERMINING COSTS OF * *..... 101
NONCONVENTIONAL SCIENTIFIC AND TECHNICAL * IN CURRENT USE **W: B3-2
SCIENTIFIC AND TECHNICAL * IN CURRENT USE **W: NONCONVENTIONAL B3-2
TECHNICAL * IN CURRENT USE **W: NONCONVENTIONAL SCIENTIFIC AND B3-2
USE **W: NONCONVENTIONAL SCIENTIFIC AND TECHNICAL * IN CURRENT B3-2
INORGANIC NOMENCLATURE
PROBLEMS *..... * IN 1966: PROGRESS AND 67
PROGRESS AND PROBLEMS *..... * IN 1966: 67
INPUT
ATOM-BY-ATOM TYPEWRITER * FOR COMPUTERIZED STORAGE AND RETRIEV* 88
CHEMICAL STRUCTURES **OR COMPUTERIZED STORAGE AND RETRIEVAL OF 88
COMPUTERIZED STORAGE AND RETRIEVAL OF CHEMICAL STRUCTURES **OR 88
RETRIEVAL OF CHEMICAL STRUCTURES **OR COMPUTERIZED STORAGE AND 88
STORAGE AND RETRIEVAL OF CHEMICAL STRUCTURES **OR COMPUTERIZED 88
STRUCTURES **OR COMPUTERIZED STORAGE AND RETRIEVAL OF CHEMICAL 88
TYPEWRITER * FOR COMPUTERIZED STORAGE AND RETRIEVAL OF CHEMICA* 88
INVERTED COORDINATE INDEX
COMPUTER PROGRAM FOR AN * ON MAGNETIC TAPE **ED TERM SEARCH: A 49
MAGNETIC TAPE **ED TERM SEARCH: A COMPUTER PROGRAM FOR AN * ON 49
PROGRAM FOR AN * ON MAGNETIC TAPE **ED TERM SEARCH: A COMPUTER 49
SEARCH: A COMPUTER PROGRAM FOR AN * ON MAGNETIC TAPE **ED TERM 49
TAPE **ED TERM SEARCH: A COMPUTER PROGRAM FOR AN * ON MAGNETIC 49
TERM SEARCH: A COMPUTER PROGRAM FOR AN * ON MAGNETIC TAPE **ED 49
WEIGHTED TERM SEARCH: A COMPUTER PROGRAM FOR AN * ON MAGNETIC * 49
ISI'S EXPERIENCE
ASCA - A SELECTIVE DISSEMINATION SYSTEM *..... * WITH 147
DISSEMINATION SYSTEM *..... * WITH ASCA - A SELECTIVE 147
SELECTIVE DISSEMINATION SYSTEM *..... * WITH ASCA - A 147
SYSTEM *..... * WITH ASCA - A SELECTIVE DISSEMINATION 147
KEYBOARDING
CHEMICAL INFORMATION *..... * 232
INFORMATION *..... * CHEMICAL 232
KINETICS
BUREAU OF STANDARDS DATA CENTERS: CHEMICAL * AND MASS SPECTRON* 31
CENTERS: CHEMICAL * AND MASS SPECTROMETRY ** OF STANDARDS DATA 31
CHEMICAL * AND MASS SPECTROMETRY ** OF STANDARDS DATA CENTERS: 31
DATA CENTERS: CHEMICAL * AND MASS SPECTROMETRY ** OF STANDARDS 31
MASS SPECTROMETRY ** OF STANDARDS DATA CENTERS: CHEMICAL * AND 31
NATIONAL BUREAU OF STANDARDS DATA CENTERS: CHEMICAL * AND MASS* 31
SPECTROMETRY ** OF STANDARDS DATA CENTERS: CHEMICAL * AND MASS 31
STANDARDS DATA CENTERS: CHEMICAL * AND MASS SPECTROMETRY ** OF 31
LANGUAGE AND MACHINES
BOOK_REVIEW: * COMPUTERS IN TRANSLATION AND LINGUISTICS *.... B 59
COMPUTERS IN TRANSLATION AND LINGUISTICS *.... BOOK_REVIEW: * B 59
LINGUISTICS *.... BOOK_REVIEW: * COMPUTERS IN TRANSLATION AND B 59
TRANSLATION AND LINGUISTICS *.... BOOK_REVIEW: * COMPUTERS IN B 59
LARGE-SCALE CHEMICAL INFORMATION
NONUNIQUE NOTATION IN A * SYSTEM *..... USE OF A 192
NOTATION IN A * SYSTEM *..... USE OF A NONUNIQUE 192
SYSTEM *..... USE OF A NONUNIQUE NOTATION IN A * 192
USE OF A NONUNIQUE NOTATION IN A * SYSTEM *..... 192
LINE NOTATIONS
COLLECTIONS **IEVAL OF ORGANIC STRUCTURES BY PERMUTED *: SMALL 85
LOW-COST STORAGE AND RETRIEVAL OF ORGANIC STRUCTURES BY PERMUT* 85
ORGANIC STRUCTURES BY PERMUTED *: SMALL COLLECTIONS **IEVAL OF 85
PERMUTED *: SMALL COLLECTIONS **IEVAL OF ORGANIC STRUCTURES BY 85
RETRIEVAL OF ORGANIC STRUCTURES BY PERMUTED *: SMALL COLLECTIO* 85
SMALL COLLECTIONS **IEVAL OF ORGANIC STRUCTURES BY PERMUTED *: 85
STORAGE AND RETRIEVAL OF ORGANIC STRUCTURES BY PERMUTED *: SMA* 85
STRUCTURES BY PERMUTED *: SMALL COLLECTIONS **IEVAL OF ORGANIC 85
LINGUISTICS
BOOK_REVIEW: INTRODUCTION TO COMPUTATIONAL * *..... B257
BOOK_REVIEW: LANGUAGE AND MACHINES COMPUTERS IN TRANSLATION A* B 59
COMPUTATIONAL * *..... BOOK_REVIEW: INTRODUCTION TO B257
COMPUTERS IN TRANSLATION AND * **EVIEW: LANGUAGE AND MACHINES B 59
INTRODUCTION TO COMPUTATIONAL * *..... BOOK_REVIEW: B257
LANGUAGE AND MACHINES COMPUTERS IN TRANSLATION AND * **EVIEW: B 59
MACHINES COMPUTERS IN TRANSLATION AND * **EVIEW: LANGUAGE AND B 59
TRANSLATION AND * **EVIEW: LANGUAGE AND MACHINES COMPUTERS IN B 59
LITERATURE
CHEMICAL AND RELATED TECHNICAL * OF WILDLIFE CONSERVATION **HE 247
CHEMICAL ENGINEERING * *..... PUBLICATION DELAYS IN THE 179
CHEMICAL PATENT * *..... DOCUMENTATION OF THE 250
CONSERVATION **HE CHEMICAL AND RELATED TECHNICAL * OF WILDLIFE 247
DELAYS IN THE CHEMICAL ENGINEERING * *..... PUBLICATION 179
DOCUMENTATION OF THE CHEMICAL PATENT * *..... 250
ENGINEERING * *..... PUBLICATION DELAYS IN THE CHEMICAL 179
PATENT * *..... DOCUMENTATION OF THE CHEMICAL 250
PUBLICATION DELAYS IN THE CHEMICAL ENGINEERING * *..... 179
RELATED TECHNICAL * OF WILDLIFE CONSERVATION **HE CHEMICAL AND 247
TECHNICAL * OF WILDLIFE CONSERVATION **HE CHEMICAL AND RELATED 247
WILDLIFE CONSERVATION **HE CHEMICAL AND RELATED TECHNICAL * OF 247
LOW-COST STORAGE
COLLECTIONS **NIC STRUCTURES BY PERMUTED LINE NOTATIONS: SMALL 85
LINE NOTATIONS: SMALL COLLECTIONS **NIC STRUCTURES BY PERMUTED 85
NOTATIONS: SMALL COLLECTIONS **NIC STRUCTURES BY PERMUTED LINE 85
ORGANIC STRUCTURES BY PERMUTED LINE NOTATIONS: SMALL COLLECTIO* 85
PERMUTED LINE NOTATIONS: SMALL COLLECTIONS **NIC STRUCTURES BY 85
RETRIEVAL OF ORGANIC STRUCTURES BY PERMUTED LINE NOTATIONS: SM* 85
SMALL COLLECTIONS **NIC STRUCTURES BY PERMUTED LINE NOTATIONS: 85
STRUCTURES BY PERMUTED LINE NOTATIONS: SMALL COLLECTIONS **NIC 85
MACHINE USE
CODE SYSTEM FOR UNIVERSAL * *..... A NUMERIC SEMANTIC 216

Figure 2. Sample page from the prototype double KWIC coordinate index

```
LIBRARY OF CONGRESS
  APPLICATIONS IN THE * SCIENCE AND TECHNOLOGY DIVISIO+  COMPUTER  63
  COMPUTER APPLICATIONS IN THE * SCIENCE AND TECHNOLOGY DIVISION=  63
  DIVISION= +OMPUTER APPLICATIONS IN THE * SCIENCE AND TECHNOLOGY  63
  SCIENCE AND TECHNOLOGY DIVISION= +OMPUTER APPLICATIONS IN THE *  63
  TECHNOLOGY DIVISION= +OMPUTER APPLICATIONS IN THE * SCIENCE AND  63
LINGUISTIC ANALYSIS
  INFORMATION RETRIEVAL= .....LINGUIA: A * SYSTEM FOR  207
  LINGUIA: A * SYSTEM FOR INFORMATION RETRIEVAL= .....  207
  RETRIEVAL= .....LINGUIA: A * SYSTEM FOR INFORMATION  207
  SYSTEM FOR INFORMATION RETRIEVAL= .....LINGUIA: A *  207
```

Figure 3. Main terms containing permuted subordinate entries derived from only a single title.

```
INDEX
    AUTHORITY LIST TO ELIMINATE SCATTERING CAUSED BY SOME SINGULA+ 277
    AUTHORITY LIST TO ELIMINATE SCATTERING CAUSED BY SOME SINGULA+ 277
    AUTOMATICALLY GENERATED AUTHORITY LIST TO ELIMINATE SCATTERIN+ 277
    AUTOMATICALLY GENERATED AUTHORITY LIST TO ELIMINATE SCATTERIN+ 277
    CAUSED BY SOME SINGULAR AND PLURAL MAIN * TERMS= +E SCATTERING 277
    CAUSED BY SOME SINGULAR AND PLURAL MAIN INDEX TERMS= +ATTERING 277
    COORDINATE *. II. USE OF AN AUTOMATICALLY GENERATED AUTHORITY+ 277
    COORDINATE INDEX. II. USE OF AN AUTOMATICALLY GENERATED AUTHO+ 277
    DOUBLE KWIC COORDINATE *. II. USE OF AN AUTOMATICALLY GENERAT+ 277
    DOUBLE KWIC COORDINATE INDEX. II. USE OF AN AUTOMATICALLY GEN+ 277
    ELIMINATE SCATTERING CAUSED BY SOME SINGULAR AND PLURAL MAIN + 277
    ELIMINATE SCATTERING CAUSED BY SOME SINGULAR AND PLURAL MAIN + 277
    GENERATED AUTHORITY LIST TO ELIMINATE SCATTERING CAUSED BY SO+ 277
    GENERATED AUTHORITY LIST TO ELIMINATE SCATTERING CAUSED BY SO+ 277
    INDEX TERMS  +ATTERING CAUSED BY SOME SINGULAR AND PLURAL MAIN 277
    INDEX. II. USE OF AN AUTOMATICALLY GENERATED AUTHORITY LIST T+ 277
    KWIC COORDINATE *. II. USE OF AN AUTOMATICALLY GENERATED AUTH+ 277
    KWIC COORDINATE INDEX. II. USE OF AN AUTOMATICALLY GENERATED + 277
    LIST TO ELIMINATE SCATTERING CAUSED BY SOME SINGULAR AND PLUR+ 277
```

Figure 4. Some permuted subordinate entries for a main term which appears more than once in a title

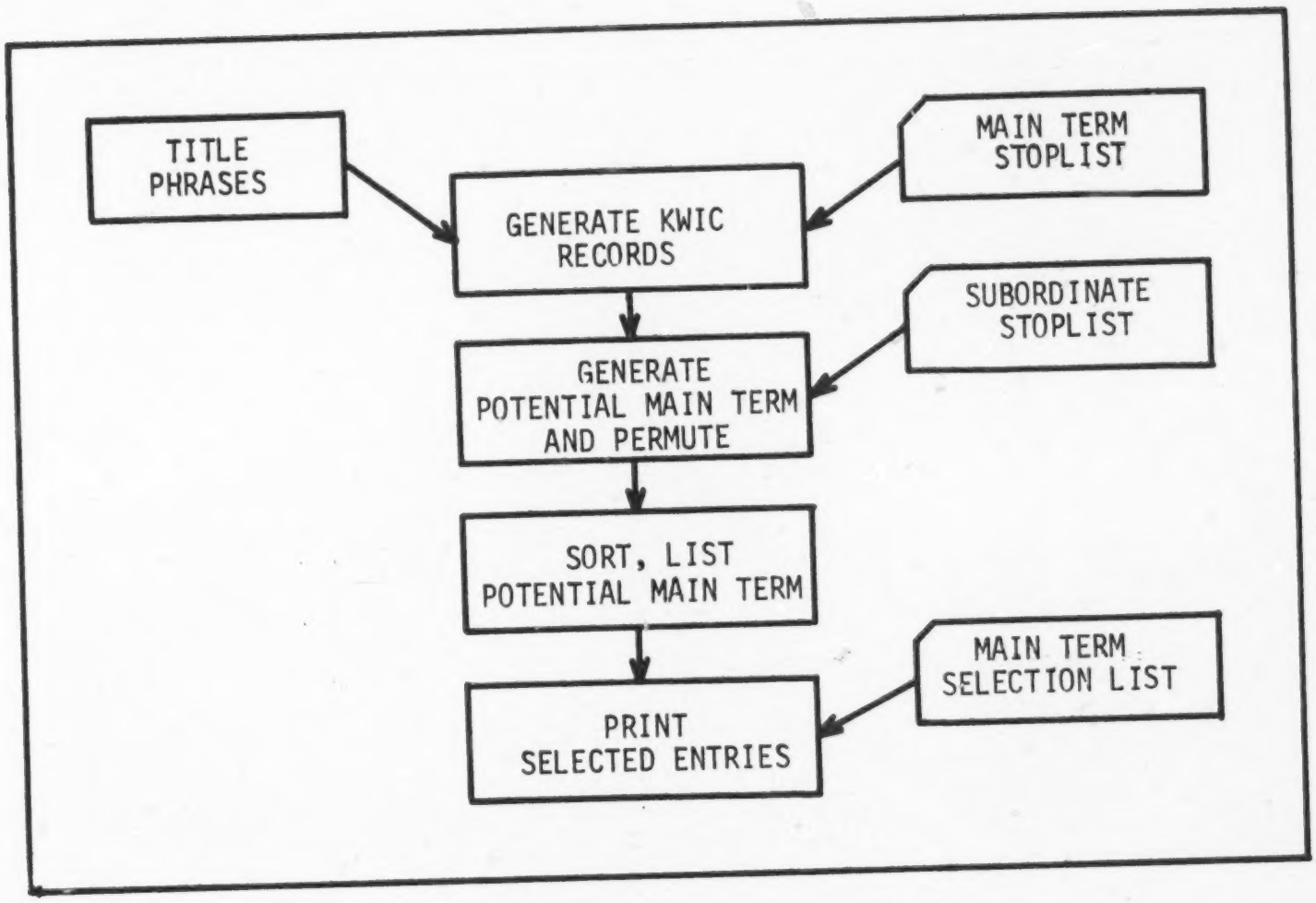

Figure 5. System design for creating the prototype double KWIC coordinate index.

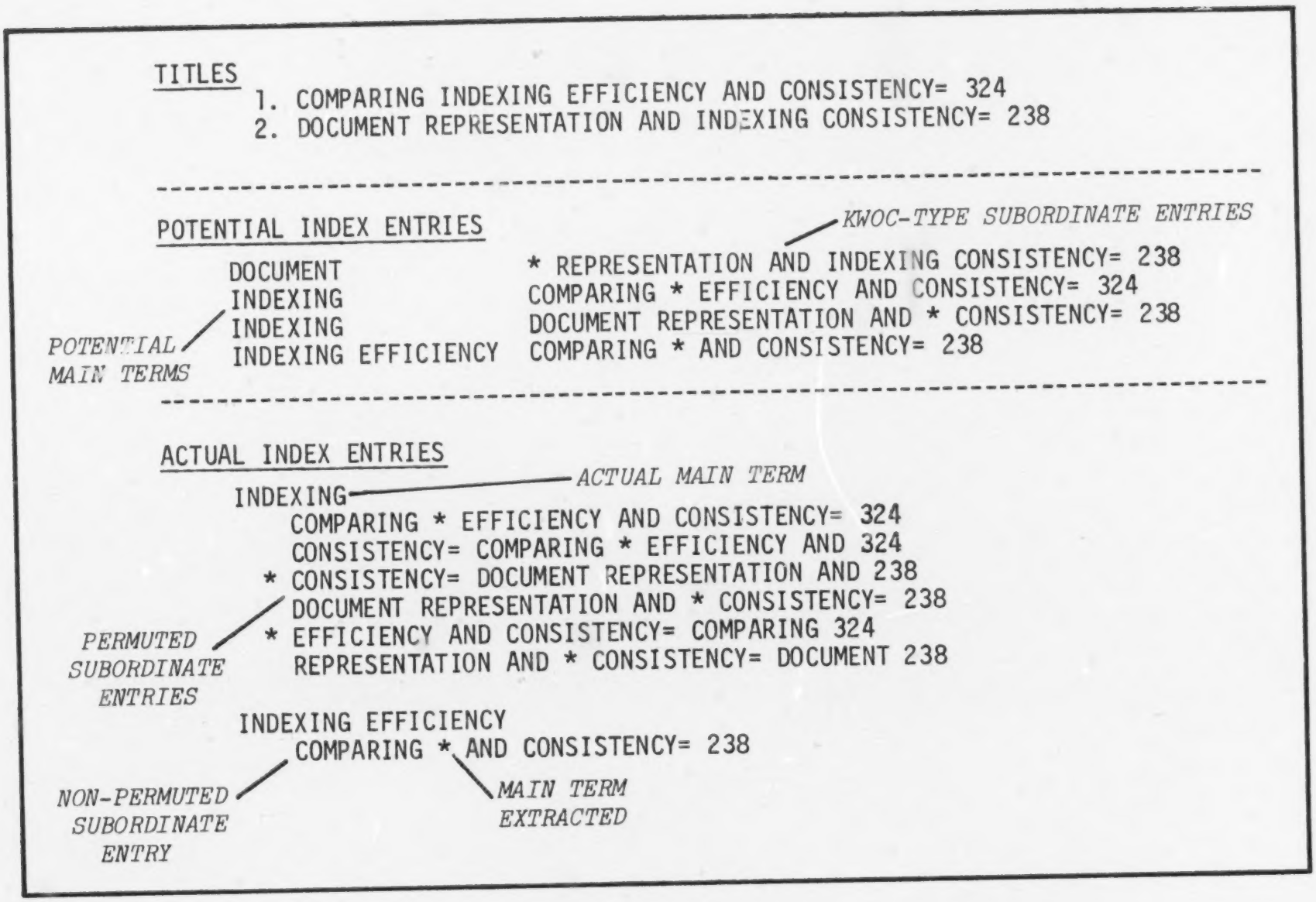
TITLES
1. COMPARING INDEXING EFFICIENCY AND CONSISTENCY= 324
2. DOCUMENT REPRESENTATION AND INDEXING CONSISTENCY= 238
POTENTIAL INDEX ENTRIES
KWOC-TYPE SUBORDINATE ENTRIES
POTENTIAL MAIN TERMS
DOCUMENT
INDEXING
INDEXING
INDEXING EFFICIENCY
* REPRESENTATION AND INDEXING CONSISTENCY= 238
COMPARING * EFFICIENCY AND CONSISTENCY= 324
DOCUMENT REPRESENTATION AND * CONSISTENCY= 238
COMPARING * AND CONSISTENCY= 238
ACTUAL INDEX ENTRIES
ACTUAL MAIN TERM
INDEXING
COMPARING * EFFICIENCY AND CONSISTENCY= 324
CONSISTENCY= COMPARING * EFFICIENCY AND 324
* CONSISTENCY= DOCUMENT REPRESENTATION AND 238
DOCUMENT REPRESENTATION AND * CONSISTENCY= 238
PERMUTED SUBORDINATE ENTRIES
* EFFICIENCY AND CONSISTENCY= COMPARING 324
REPRESENTATION AND * CONSISTENCY= DOCUMENT 238
INDEXING EFFICIENCY
COMPARING * AND CONSISTENCY= 238
NON-PERMUTED SUBORDINATE ENTRY
MAIN TERM EXTRACTED

Figure 6. Construction of index entries for the modified double KWIC coordinate index.

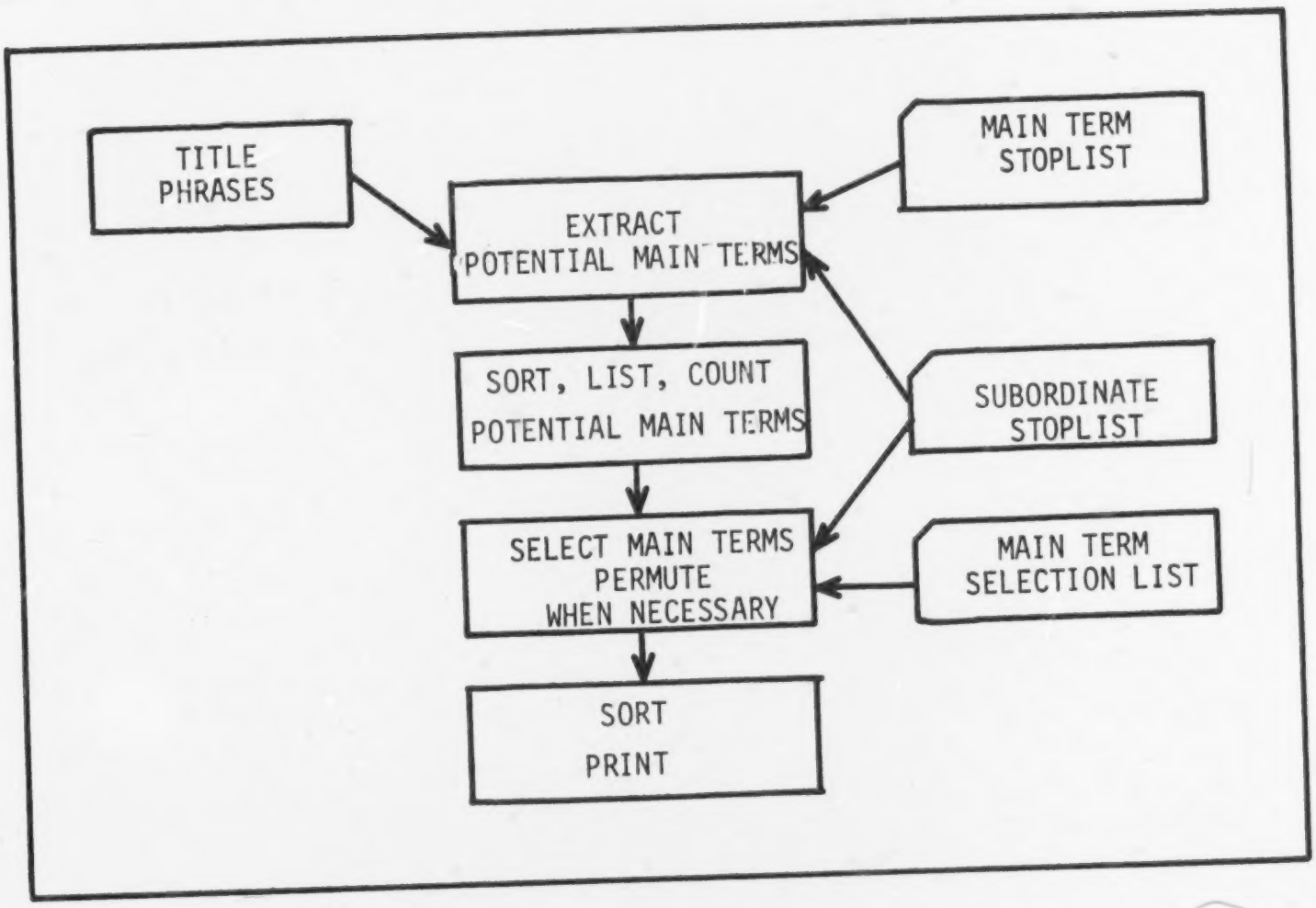

Figure 7. System design for creating the modified double KWIC coordinate index.

WORD DELIMITERS

TERMINAL ".,;:?()"
NONTERMINAL "Ø-/"

TITLE

DASAR: A COMPUTER-BASED DATA STORAGE AND DATA RETRIEVAL SYSTEM= 245

POTENTIAL MAIN TERMS

DASAR
COMPUTER
DATA
STORAGE
RETRIEVAL

COMPUTER BASED
DATA STORAGE
DATA RETRIEVAL
RETRIEVAL SYSTEM

COMPUTER BASED DATA
STORAGE AND DATA
DATA RETRIEVAL SYSTEM

SOME POTENTIAL INDEX ENTRIES

DATA DASAR: A COMPUTER-BASED * STORAGE AND * RETRIEVAL SYSTEM= 245
DATA STORAGE DASAR: A COMPUTER-BASED * AND DATA RETRIEVAL SYSTEM= 245

Figure 8. Potential main terms and potential index entries generated from a title.

SEQ #	FREQ	POTENTIAL MAIN TERM
29	13	DATA
30	1	DATA AQUISITION
31	3	DATA BASES
32	2	DATA RETRIEVAL
33	2	DATA RETRIEVAL SYSTEM
34	4	DATA STORAGE

Figure 9. Portion of a printed list used for selection of actual main terms

INDEX
THE DOUBLE-KWIC COORDINATE * II. USE OF AN AUTOMATICALLY
GENERATED AUTHORITY LIST TO ELIMINATE SCATTERING CAUSED BY
SOME SINGULAR AND PLURAL MAIN * TERMS=.....277

INDEXING
COMPARING * EFFICIENCY AND CONSISTENCY=..... 324
CONSISTENCY=.....COMPARING * EFFICIENCY AND 324
* CONSISTENCY=.....DOCUMENT REPRESENTATION AND 238
DOCUMENT REPRESENTATION AND * CONSISTENCY=..... 238
* EFFICIENCY AND CONSISTENCY=.....COMPARING 324
REPRESENTATION AND * CONSISTENCY=.....DOCUMENT 238

Figure 10. Examples of two types of subordinate entries found in a modified double KWIC coordinate index.

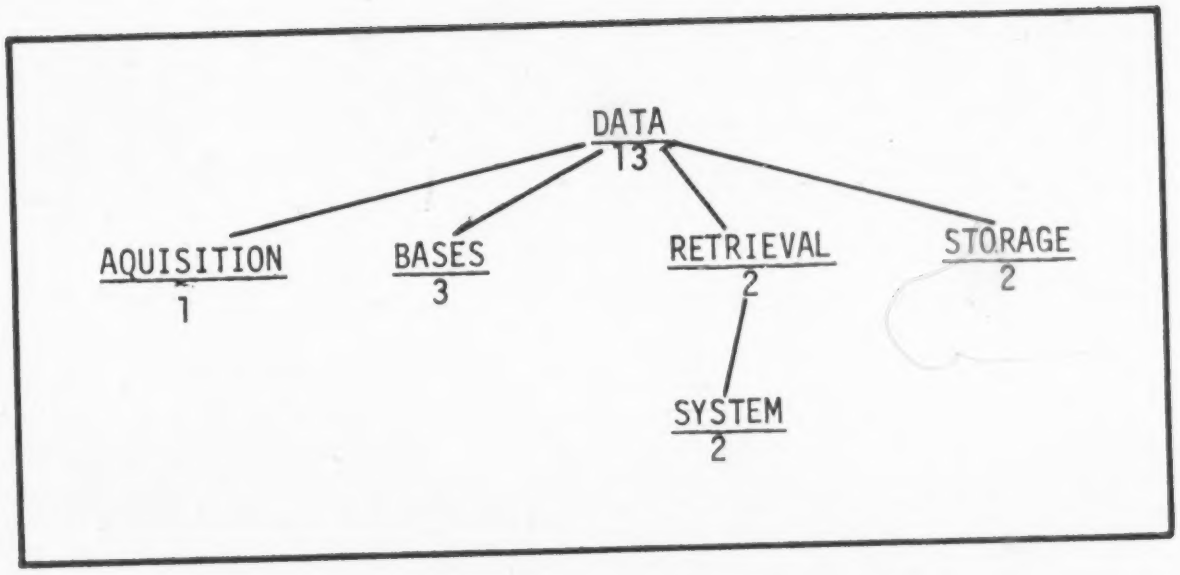

Figure 11. Tree representation of the potential-main-term list